

Warcbase:

Using Scalable Web Analytics to Analyze Canadian Collections En Masse

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Nick Ruest
Digital Assets Librarian
@ruebot



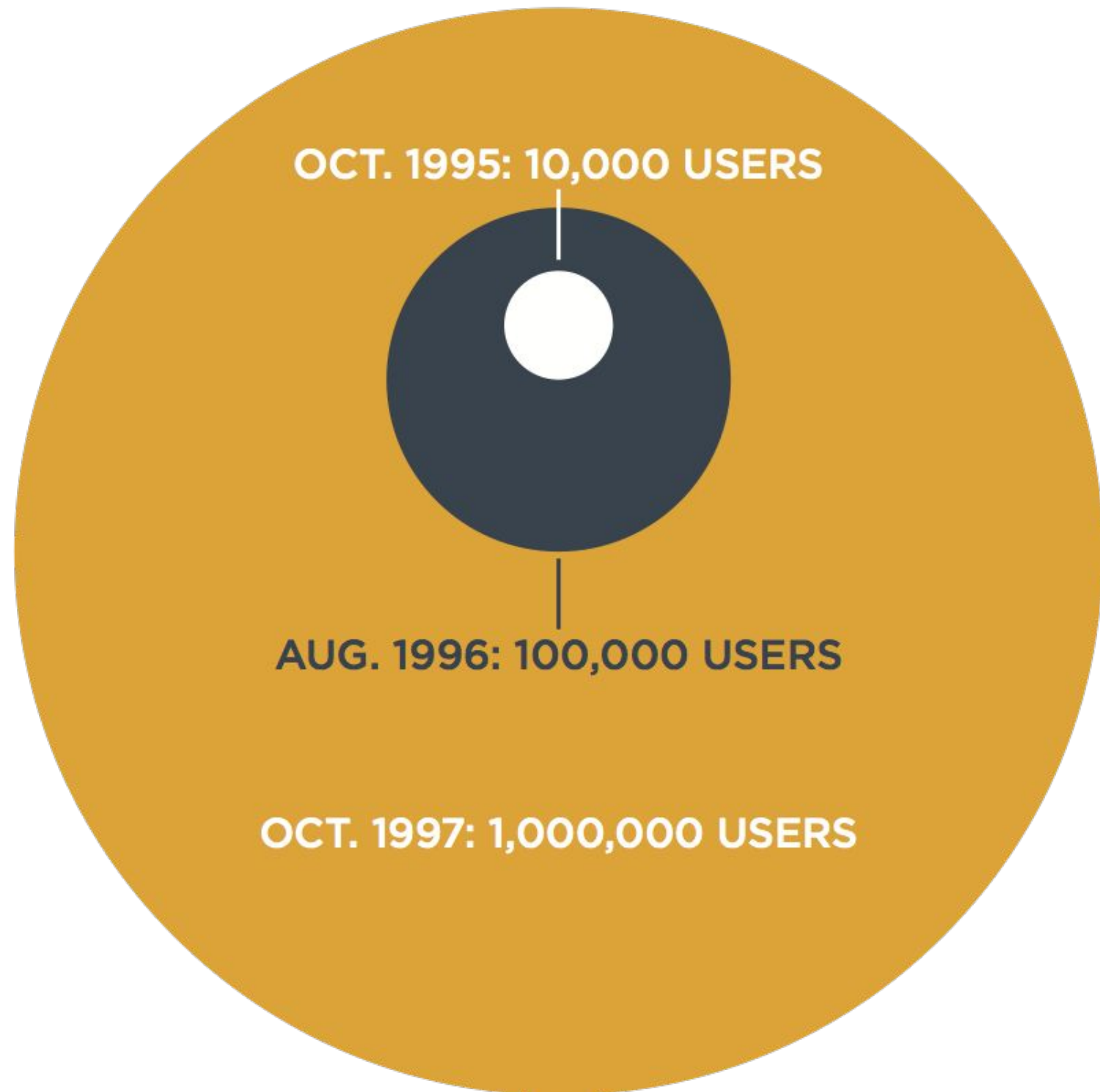
The Web as a Primary Source

- **Web archives will fundamentally affect the way historians write history**
- We will have easier access to information on a previously-unknown scale, as well as improved capability to parse it;
- This information will be left by people who rarely before left historical sources;

**We live and experience our
lives online, and historians
need to study this.**

And it's all
happening on a
scale
historians
have difficulty
imagining!

GEOCITIES USERS:



Scarcity



The background of the image shows a series of grey archival boxes hanging from a metal rack. Each box has a white label with black text. The labels are organized into columns and rows. Some labels are partially obscured by the boxes in front of them. The text on the labels includes "CONGRESSIONAL ARCHIVES", "THOMAS P. O'NEILL PAPERS", "PARTY LEADERSHIP / ADMINISTRATIVE FILES", "Democratic Party", "BOX 29", "BOX 30", "JOHN J. BURNS LIBRARY", and "BOSTON COLLEGE". There are also handwritten notations like "Series V" and "Subseries F". The overall scene suggests a large, organized collection of historical documents.

Scarcity

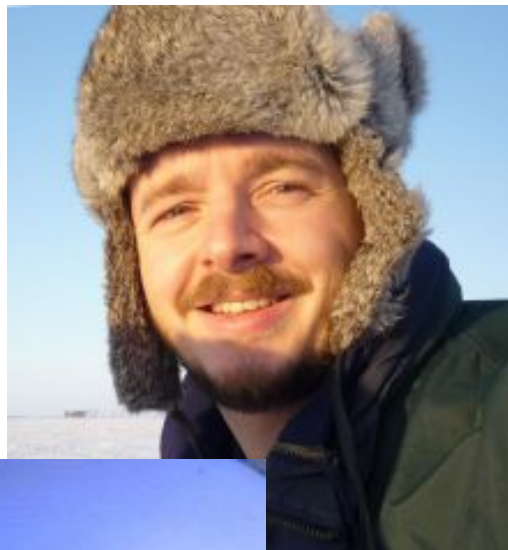
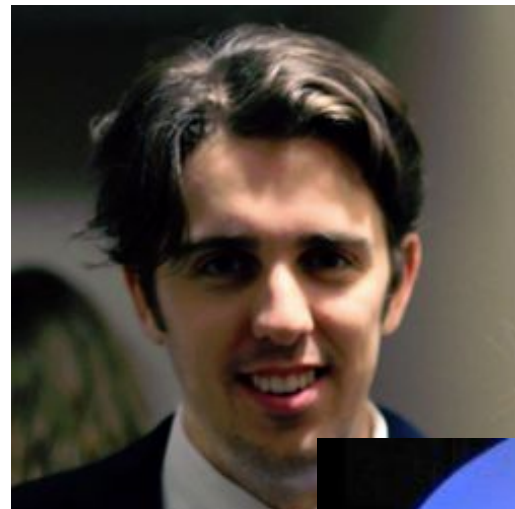
Abundance

**Historians and other
humanists can't do it
alone**

**We need
collaboration!**

Web Archives for Historical Research

Historians



Computer Scientists



Librarians

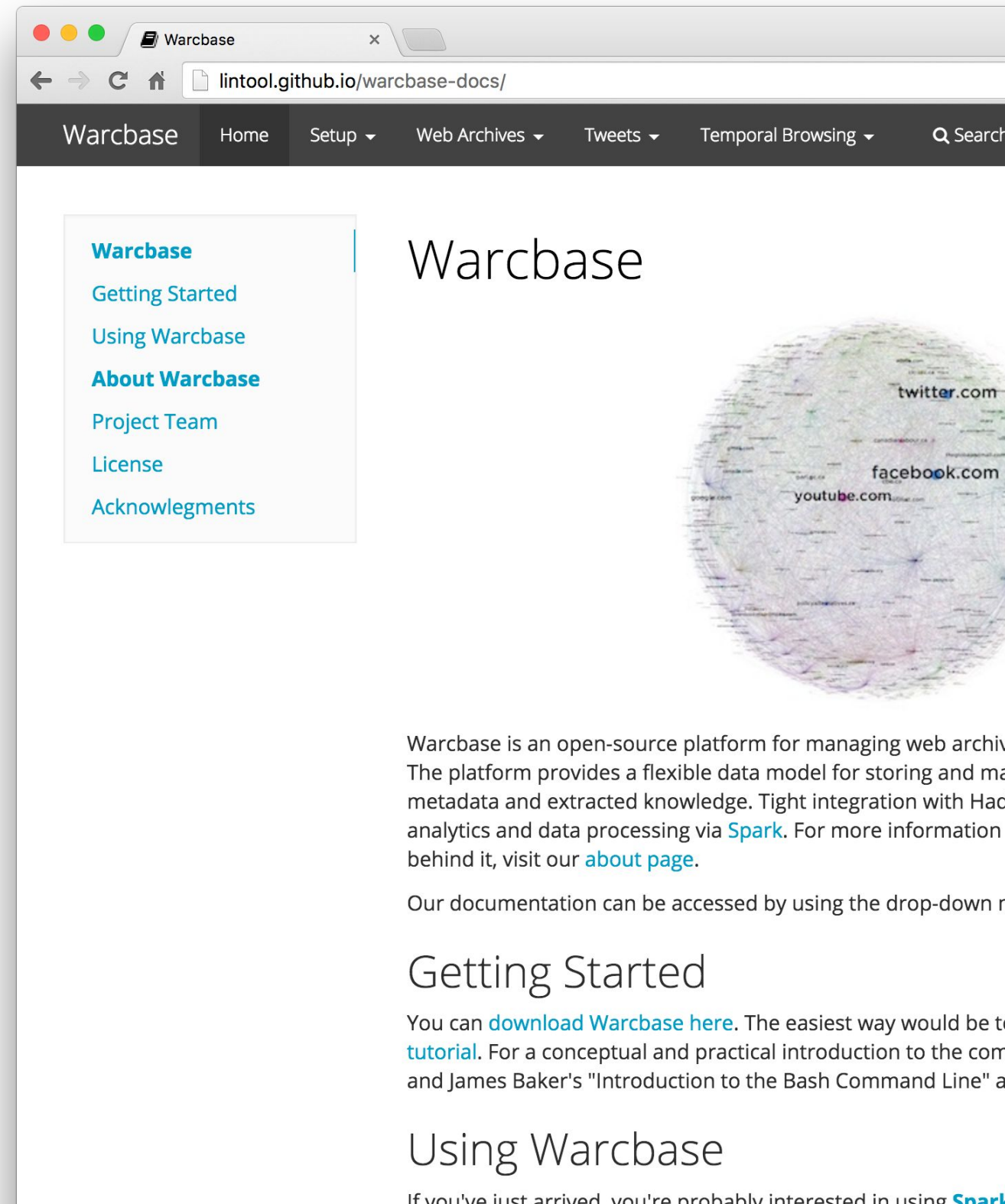


Governance



Warcbase

- **Jimmy Lin** (main developer, CS/lead), **Ian Milligan** (co-lead, history), **Jeremy Wiebe** (history/PhD), **Alice Zhou** (computer science, undergrad), **Youngbin Kim** (computer science, undergrad), **Nick Ruest** (librarian @ York)
- Currently using it on the **GeoCities** and **Canadian Politics** web archives, as well as WALK (61 Archive-It collections, 6 institutions)



docs.warcbase.org

The screenshot shows a web browser window with the URL `lintool.github.io/warcbase-docs/Spark-Extracting-Domain-Level-Plain-Text/`. The page title is "Extracting Domain Level Plain Text". On the left, there is a sidebar with a list of links: "Extracting Domain Level Plain Text" (highlighted), "All plain text", "Plain text by domain", "Plain text by URL pattern", "Plain text minus boilerplate", "Plain text filtered by date", "Plain text filtered by language", and "Plain text filtered by keyword". The main content area has the title "Extracting Domain Level Plain Text" and a subheading "All plain text". Below this, a paragraph states: "This script extracts the crawl date, domain, URL, and plain text from HTML files in the sample ARC data (and saves the output to out/)." A code block contains the following Scala code:

```
import org.warcbase.spark.rdd.RecordRDD._
import org.warcbase.spark.matchbox.{RemoveHTML, RecordLoader}

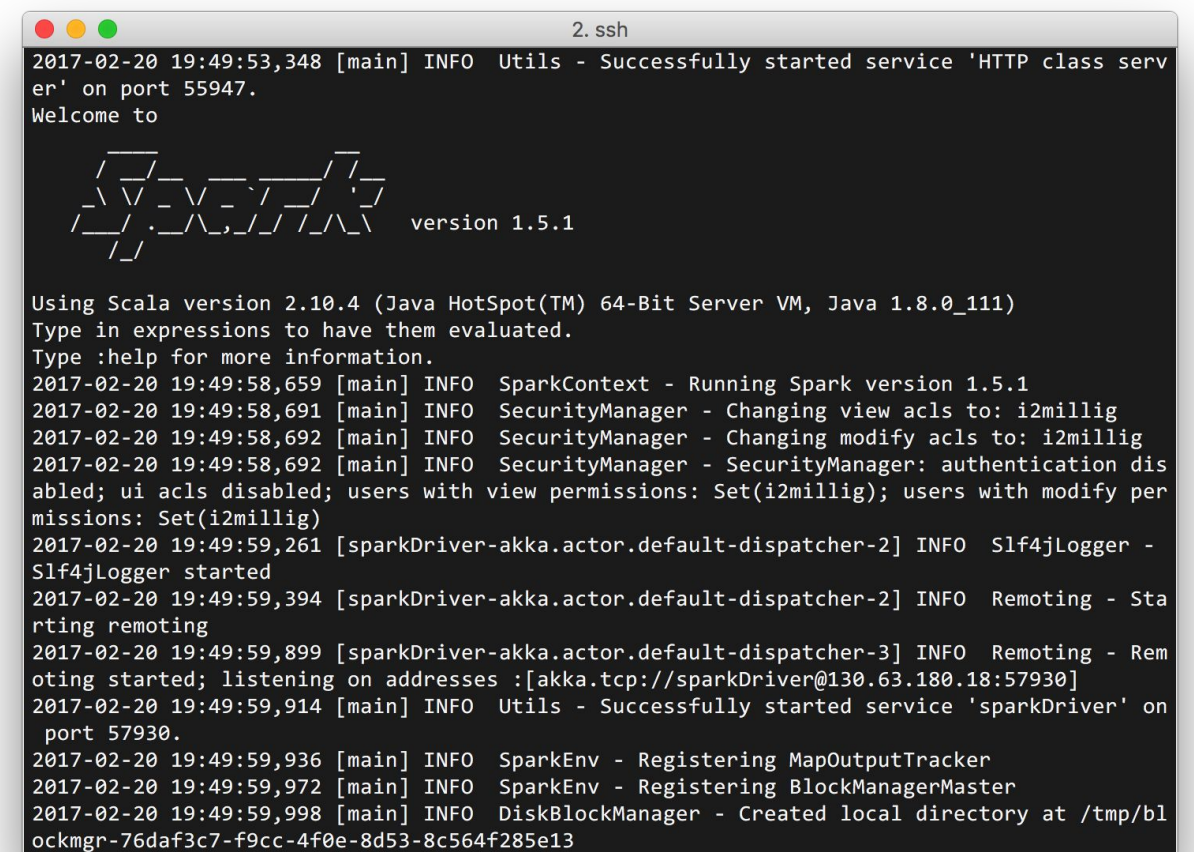
RecordLoader.loadArchives("src/test/resources/arc/example.arc.gz", sc)
  .keepValidPages()
  .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getContentString)))
  .saveAsTextFile("out/")
```

Below the code block, a paragraph says: "If you wanted to use it on your own collection, you would change 'src/test/resources/arc/example.arc.gz' to the directory with your own ARC or WARC files, and change 'out/' on the last line to where you want to save your output data." Another paragraph notes: "Note that this will create a new directory to store the output, which cannot already exist." A final paragraph states: "If you want to run it in your Spark Notebook, the following script will show in-notebook plain text:" followed by a code block with the following Scala code:

```
val r = RecordLoader.loadArchives("/path/to/warcs", sc)
  .keepValidPages()
  .map(r => {
```

Installing Warchbase

- Install Spark, clone warchbase, build with maven
- Launch Spark Shell
- Run simple scripts (begin by copy-and-replacing our existing ones to get a feel for the interface)
- In the future, hoping to move to PySpark



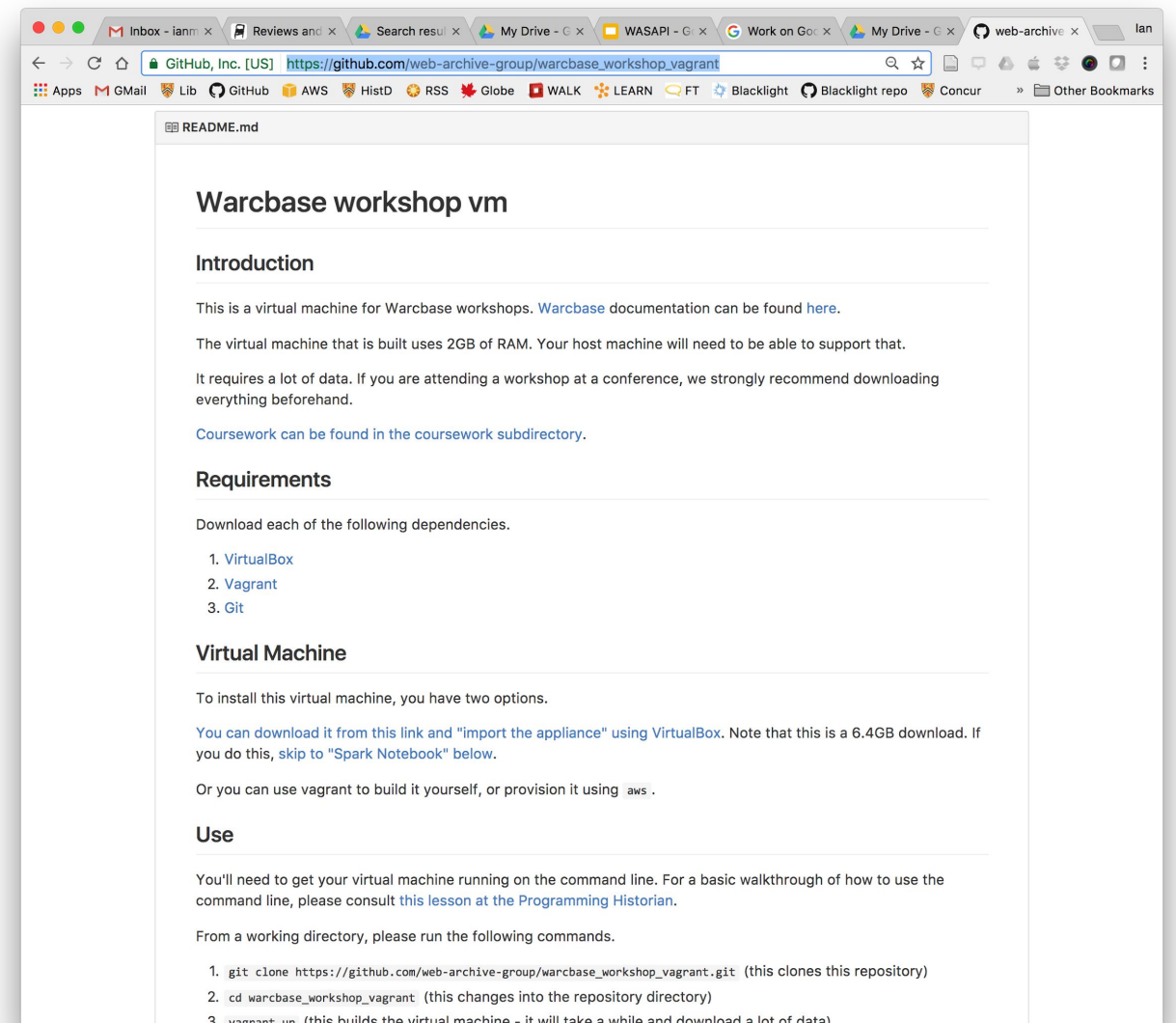
```
2. ssh
2017-02-20 19:49:53,348 [main] INFO  Utils - Successfully started service 'HTTP class serv
er' on port 55947.
Welcome to

  _ _ _ _ _
 / _ _ _ _ \
( _ _ _ _ )
 \ _ _ _ _ /
  _ _ _ _ _
  version 1.5.1

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_111)
Type in expressions to have them evaluated.
Type :help for more information.
2017-02-20 19:49:58,659 [main] INFO  SparkContext - Running Spark version 1.5.1
2017-02-20 19:49:58,691 [main] INFO  SecurityManager - Changing view acls to: i2millig
2017-02-20 19:49:58,692 [main] INFO  SecurityManager - Changing modify acls to: i2millig
2017-02-20 19:49:58,692 [main] INFO  SecurityManager - SecurityManager: authentication dis
abled; ui acls disabled; users with view permissions: Set(i2millig); users with modify per
missions: Set(i2millig)
2017-02-20 19:49:59,261 [sparkDriver-akka.actor.default-dispatcher-2] INFO  Slf4jLogger -
Slf4jLogger started
2017-02-20 19:49:59,394 [sparkDriver-akka.actor.default-dispatcher-2] INFO  Remoting - Sta
rting remoting
2017-02-20 19:49:59,899 [sparkDriver-akka.actor.default-dispatcher-3] INFO  Remoting - Rem
oting started; listening on addresses :[akka.tcp://sparkDriver@130.63.180.18:57930]
2017-02-20 19:49:59,914 [main] INFO  Utils - Successfully started service 'sparkDriver' on
port 57930.
2017-02-20 19:49:59,936 [main] INFO  SparkEnv - Registering MapOutputTracker
2017-02-20 19:49:59,972 [main] INFO  SparkEnv - Registering BlockManagerMaster
2017-02-20 19:49:59,998 [main] INFO  DiskBlockManager - Created local directory at /tmp/bl
ockmgr-76daf3c7-f9cc-4f0e-8d53-8c564f285e13
```


.... Or the Warcbase Vagrant machine!

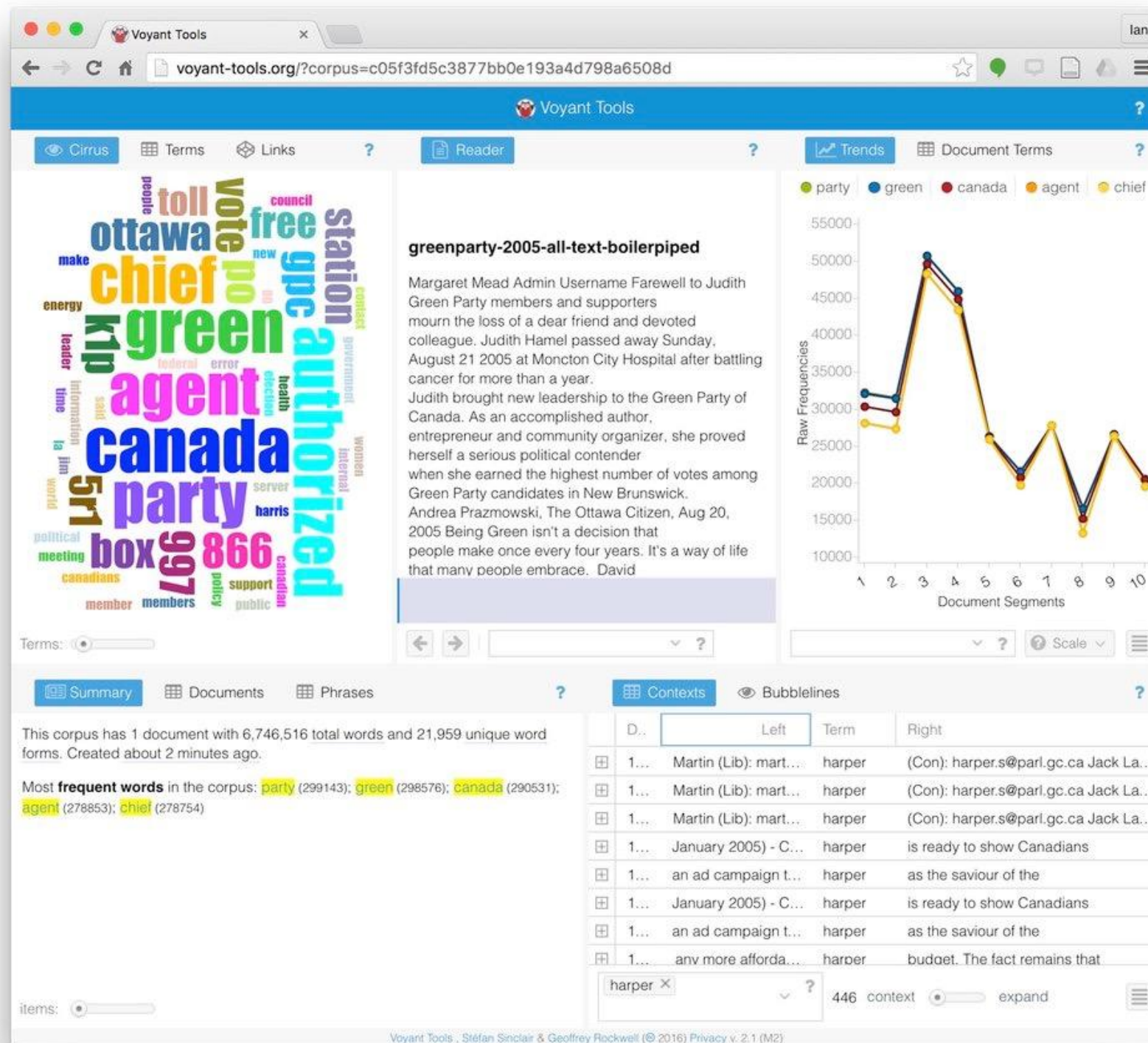
- Digital pedagogy FTW
- https://github.com/web-archive-group/warcbase_workshop_vagrant



Extract all Text

```
1. i2millig@rho: ~/derivatives/cpp.all.plaintext (ssh)
Python  bash  bash  bash  i2millig@rho:...  i2millig@rho:...  bash
(20060222,liberal.ca,http://liberal.ca/bio_e.aspx?&id=35049,Liberal Party of Canada HOME THE TEAM THE P
ARTY MEDIA CENTRE COMMISSIONS YOUR RIDING Omar Alghabra www.omaralghabra.com Home > Mississauga--E
rindale Riding Map (PDF) Omar Alghabra came to Canada at a very young age, and immediately knew Canada
was his home. He is was first elected 2006 as the Member of Parliament for Mississauga-Erindale. Mr. Al
ghabra is an experienced entrepreneur. For the past six years, he has worked for a large multinational
corporation, carrying out different responsibilities including quality assurance, project management, s
ales, contract management and management of a complete department handling a global mandate. Mr. Alghab
ra is an active member of his community. He is the former National President of the Canadian Arab Feder
ation (2004-2005) and a former member of the Community Editorial Board for the Toronto Star. (2003-2004
). Mr. Alghabra is currently a member of the Diversity Council for General Electric Canada and is activ
e in Junior Achievement for the Toronto Region. He was a member of the Multicultural Inter-Agency of Pe
el from 2001 to 2002. Mr. Alghabra has a degree in Mechanical Engineering from Ryerson University and a
Masters in Business Administration (MBA) from York University. Omar Alghabra 790 Burnamthorpe West,
Unit 10 905-276-2806 info@omaralghabra.ca Riding President Elias Hazineh Send an email
Home | News | Your Riding | Contact Us | français This website is the property of the
Liberal Party of Canada and may not be reproduced in whole or in part without express written permissio
n. © Liberal Party of Canada 2006. All rights reserved. Authorized by the registered agent for the Libe
ral Party of Canada. Privacy Policy)
(20060222,liberal.ca,https://liberal.ca/news_e.aspx?id=11470,Liberal.ca HOME THE TEAM THE PARTY MEDIA C
ENTRE COMMISSIONS YOUR RIDING Celebrating our National Flag February 15, 2006 February 15 is Nationa
l Flag Day in Canada, which marks the 41st anniversary of the first raising of the maple leaf flag on P
arliament Hill. Today is a celebration of our shared values, common citizenship and sense of pride in t
his great country we call home. The Canadian flag is one of the most recognizable symbols in the world
and flies proudly to remind us all of who we are and where we come from. The maple leaf's symbolic orig
ins date back to the beginning of our nation's history, while the red and white bars on the flag repres
ent strength and unity. Canada's flag was adopted in 1964 under the courageous leadership of Liberal Pr
ime Minister Lester B. Pearson. The idea of changing the Red Ensign which featured the Britain's Union
Jack, was very controversial at the time, with the Conservative Party strongly opposed to changing the
status quo. Facing strong Conservative resistance in the House of Commons, Pearson's minority governme
nt fought hard in the name of national unity and Canada's multicultural future to make the new flag a r
eality. In an impassioned speech to the House of Commons, Pearson said: "Mr. Speaker, it is for this g
eneration, for this Parliament, to give them and to give us all a common flag; a Canadian flag which, w
hile bringing together but rising above the landmarks and milestones of the past, will say proudly to t
he world and to the future: I stand for Canada." Thanks to Pearson's courageous leadership, Canadians a
cross this great nation celebrate our flag and what it stands for – a country and a citizenship that ar
e the envy of the world. Home | News | Your Riding | Contact Us | fran
cais This website is the property of the Liberal Party of Canada and may not be reproduced in whole or
```


Extract all Text



Extract Entities

The word cloud displays names and their frequency of mentions across five years. The names are arranged in a grid-like structure, with larger text indicating more mentions. A blue box highlights 'Elizabeth May' with '10 mentions'.

Year	Names and Mentions
200606	Andrew Lewis, Bill, Bill Hulet, Brown, Bruce Abel, Bush, Camille Labchuk, Chandler, Cherfi, Chernushenko, David, David Chernushenko, David Chernushenko, David Kay, Derek Pinto, Ed Broadbent, Elizabeth May, Eric Walton, Fannon, Gomery, Green, Harper, Harris, Jim, Jim Fannon, Jim Harris, Jim Harris Speech, John, Julie Baribeau, Junker, Kevin Colton, Labchuk, Layton, Leonardo DiCaprio, Manley, Mark Brooks, Mark MacGillivray, Martin, Michael Robinson, Milliken, Paul Martin, Paul Martin
200607	Adrianne Carr, Andrew Lewis, Bill, Bill Hulet, Brown, Bruce Abel, Camille Labchuk, Chandler, Cherfi, Chernushenko, David, David Chernushenko, David Chernushenko, David Kay, Derek Pinto, Dietrich, Ed Broadbent, Elizabeth May, Eric Walton, Fannon, Gomery, Green, Harper, Harris, Jim, Jim Fannon, Jim Harris, Jim Harris Speech, John, Julie Baribeau, Junker, Kevin Colton, Labchuk, Layton, Manley
200608	Adrianne Carr, Allan Gribbin, Amélie Gingras, Andrew Lewis, Bill, Bill Hulet, Brown, Bruce Abel, Bush, Chandler, Cherfi, Chernushenko, Clements Verhoeven, David, David Kay, Derek Pinto, Dietrich, Ed Broadbent, Elizabeth May, Eric Walton, Fannon, Gomery, Green, Harper, Harris, Jim, Jim Harris, Jim Harris Speech, John, Junker, Kevin Colton, Kootenay-Columbia Jo..., Labchuk, Layton, Lawrence Redfern, Manley, Mark Brooks
200609	Adrianne Carr, Amélie Gingras, Brown, Bruce Abel, Bush, Cameron Wigmore, Chandler, Cherfi, Chernushenko, Chretien, David, David Chernushenko, David Kay, Derek Pinto, Dietrich, Dion, Elizabeth, Elizabeth May, Elizabeth May Say, Eric Walton, Gagnon, Gomery, Green, Grenon, Halton, Harper, Harris, Jim, Jim Harris, Jim Harris Speech, John, Labchuk, Loughheed, Mackenzie, Manley, Martin, May, Mona Elaine Adilman ..., Paul Martin, Peter Foster, Pierre Pettigrew, Schiller
200610	Ambrose, Andrew Lewis, Bill, Bill Clinton, Bush, Chandler, Cherfi, Chernushenko, Chris Alders, Daphne Wysham, David, David Chernushenko, David Cox, David Kay, David Suzuki, Derek, Derek Pinto, Dundas, Elizabeth, Elizabeth Goes, Elizabeth May, Elizabeth May Say, Eric Walton, Gagnon, Gomery, Green, Grenon, Halton, Harper, Harris, Jim, Jim Harris, John, Jude Larkin, Judith, Kyle Grice, Labchuk, Manley, Mark MacGillivray, Martin, May, Melanie Ransom, Michael Grayson, Michele, Paul Martin, Paul Martin, Richard Reble, Sharon Labchuk, Sharon Labchuk

Extract Entities

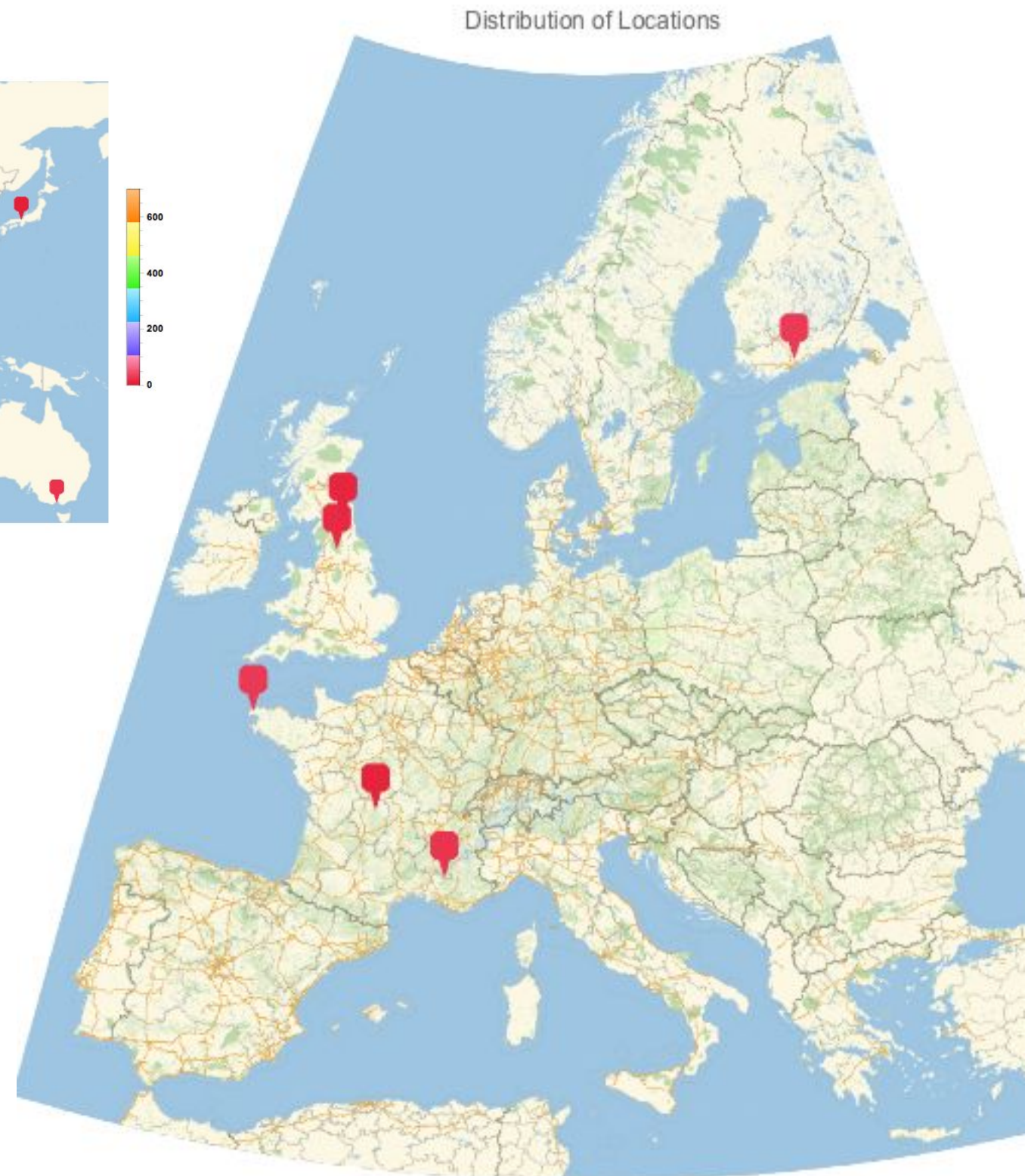


Extract Entities



```
In[95]:= int = SemanticInterpretation[#] & /@processedfreq[[All, 1]]
```

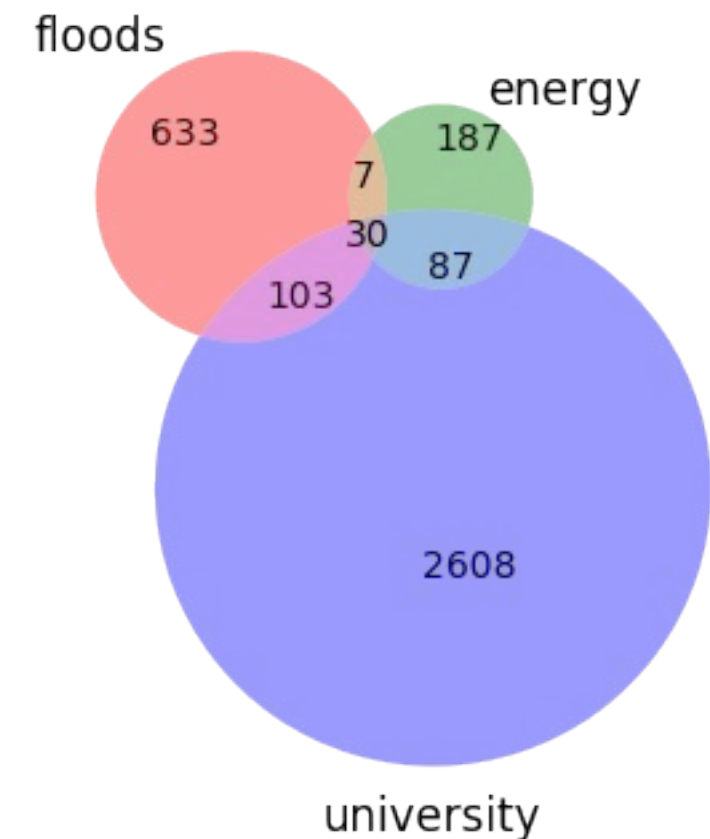
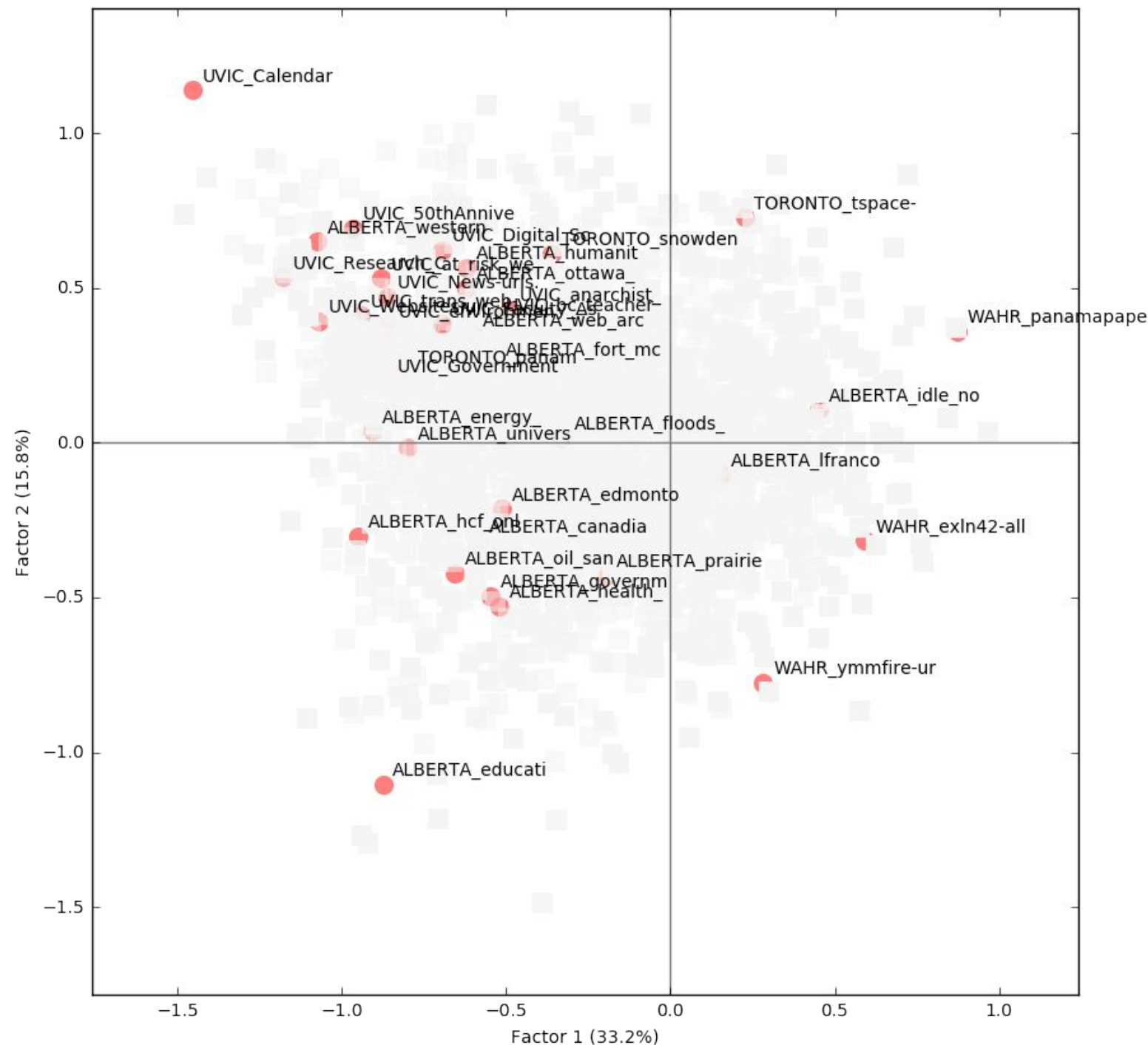
Out[95]= { Canada , Calgary , Colombia , Montreal , \$Failed , Ontario, Canada , Afghanistan ,
Ottawa , Manitoba, Canada , British Columbia, Canada , Toronto , Nova Scotia, Canada ,
Saskatchewan, Canada , Quebec, Canada , Alberta, Canada , Winnipeg , Washington ,
Canada , Nunavut, Canada , White Horse , United States , Petawawa , Mumbai ,
Lima , The Americas , Saskatoon , Peru , Edmonton , Mexico , Chicoutimi-Jonquiere ,
New Brunswick, Canada , United States , Newfoundland and Labrador, Canada , Qandahar ,
\$Failed , Asia-Pacific , Newfoundland and Labrador, Canada , Victoria , United States ,
Quebec City , India , \$Failed , Atlantic Ocean , St. Germain , Durham , Battle of Waterloo ,



Finding Sites of Interest



Exploring collection coverage and curatorial models



So we have this
tool... how have we
used it?

Scaling up



Web Archives for Longitudinal Knowledge

- **Ian Milligan** (Co-PI, UW) + **Nick Ruest** (Co-PI, York), w/ **Geoff Harder**, **Todd Suomela**, **Sonya Betz**, **Peter Binkley**, **Geoffrey Rockwell**, **Umar Qasim** (Alberta), **Jefferson Bailey** (Internet Archive), and **John Simpson** (Compute Canada).



Web Archives for Longitudinal Knowledge

- 6 universities (SFU, Toronto, Alberta, Victoria, Winnipeg, and Dalhousie)
- 61 collections
- 16 TB of web archival collections
- Ingesting them all to generate warcbase derivatives and searchable index



Web Archives for Longitudinal Knowledge

- Mass Ingest Script to take each collection and:
 - Extract hyperlinks and generate gephi files;
 - Extract all URLs and domain counts;
 - Extract plain text;

```
2. ssh
ubuntu@compute-canada-will-accidentally-delete-this:~/production$ cat template.scala
import org.warchbase.spark.matchbox._
import org.warchbase.spark.rdd.RecordRDD._
import org.warchbase.spark.matchbox.{RemoveHTML, RecordLoader, ExtractBoilerpipeText}
val ${COLLECTION} = RecordLoader.loadArchives("/data/${COLLECTION}/*.gz", sc).keepValidPages().map(r => (r.getCrawlMonth, ExtractDomain(r.getUrl))).countItems().saveAsTextFile("/data/derivatives/urls/${COLLECTION}")
RecordLoader.loadArchives("/data/${COLLECTION}/*.gz", sc).keepValidPages().map(r => (r.getCrawlDate, ExtractLinks(r.getUrl, r.getContentString))).flatMap(r => r._2.map(f => (r._1, ExtractDomain(f._1.replaceAll("^\\s*www\\.\"", ""), ExtractDomain(f._2.replaceAll("^\\s*www\\.\"", "")))).filter(r => r._2 != "" && r._3 != "").countItems().filter(r => r._2 > 5).saveAsTextFile("/data/derivatives/links/${COLLECTION}")
val ${COLLECTION}gephi = RecordLoader.loadArchives("/data/${COLLECTION}/*.gz", sc).keepValidPages().map(r => (r.getCrawlDate, ExtractLinks(r.getUrl, r.getContentString))).flatMap(r => r._2.map(f => (r._1, ExtractDomain(f._1.replaceAll("^\\s*www\\.\"", ""), ExtractDomain(f._2.replaceAll("^\\s*www\\.\"", "")))).filter(r => r._2 != "" && r._3 != "").countItems().filter(r => r._2 > 5)
WriteGDF(${COLLECTION}gephi, "/data/derivatives/gephi/${COLLECTION}.gdf")
RecordLoader.loadArchives("/data/${COLLECTION}/*.gz", sc).keepValidPages().map(r => (r.getCrawlMonth, r.getDomain, r.getUrl, ExtractBoilerpipeText(r.getContentString))).saveAsTextFile("/data/derivatives/text/${COLLECTION}")
exit

ubuntu@compute-canada-will-accidentally-delete-this:~/production$
```

WebArchives.ca (Blacklight)

- Warcbase with a GUI front end
- Archivists/Librarians bring their collections, generate derivatives
- Hosted with Compute Canada

Blacklight Search Results

localhost:3000/?utf8=✓&search_field=nce

WALK

All Fields

nce

Search

Limit your search

General Content Type

html	4,240
pdf	783
text	295
other	209
audio	0
image	0
powerpoint	0
video	0
word	0

Domain

Links Domains

Institution

University of Alberta Libraries	4,671
University of Victoria Libraries	457
University of Toronto Libraries	199
Simon Fraser University Library	90
Dalhousie University Libraries	62
University of Winnipeg Library	48

Collection Name

Prairie Provinces	3,940
Alberta Oil Sands	153
University of Alberta Websites	141
University of Victoria Research Centres, Groups, and Corporate Entities	130
University of Victoria Websites	129
Idle No More	92
Public Knowledge Project	87
University of Victoria Digital Scholarship Websites	80
T-Space	77
British Columbia Lo-	76

You searched for: nce

Start Over

« Previous | 1 - 10 of 5,527 | Next »

10 per page -

1. APPLICATION: bcgenesis.uvic.ca [crawled on: 2014-08-18] [Bookmark](#)

Host: bcgenesis.uvic.ca

Crawl Date: 2014-08-18T01:28:40Z

Content Type: application/xml

Domain: uvic.ca

Institution: University of Victoria Libraries

Collection Name: University of Victoria Digital Scholarship Websites

Collection Number: 4376

2. APPLICATION: mapoflondon.uvic.ca [crawled on: 2015-10-02] [Bookmark](#)

Host: mapoflondon.uvic.ca

Crawl Date: 2015-10-02T04:10:09Z

Content Type: application/xml

Domain: uvic.ca

Institution: University of Victoria Libraries

Collection Name: University of Victoria Digital Scholarship Websites

Collection Number: 4376

3. APPLICATION: youtube.com [crawled on: 2015-02-15] [Bookmark](#)

Host: youtube.com

Crawl Date: 2015-02-15T21:10:47Z

Content Type: application/xml

Domain: youtube.com

Institution: University of Winnipeg Library

Collection Name: University of Winnipeg Websites

Collection Number: 3832

4. APPLICATION: youtube.com [crawled on: 2015-11-02] [Bookmark](#)

Host: youtube.com

Crawl Date: 2015-11-02T05:40:32Z

Content Type: application/xml

Domain: youtube.com

Institution: University of Toronto Libraries

Collection Name: T-Space

Collection Number: 6473

5. APPLICATION: youtube.com [crawled on: 2015-02-17] [Bookmark](#)

Host: youtube.com

Crawl Date: 2015-02-17T13:37:32Z

Content Type: application/xml

Domain: youtube.com

Institution: University of Winnipeg Library

Where are we at now?

Getting the data;

*Definitely a need for a
standardized
interaction model!*

UKWA
and
Andy Jackson,
Thank you!

Indexing

```
2017-02-21 21:46:27 ERROR HTMLAnalyser:134 - Failed to canonicalise host: French Language Summer Camps Level I: org.apache.commons.httpclient.URIException: gnu.inet.encoding.IDNAException: Contains non-LDH characters. french%20language%20summer%20camps%20level%20i
2017-02-21 22:07:45 ERROR HTMLAnalyser:134 - Failed to canonicalise host: josephmoreau.schoolappointments.
com: org.apache.commons.httpclient.URIException: gnu.inet.encoding.IDNAException: Contains non-LDH characters. josephmoreau.schoolappointments.%20com
2017-02-21 22:07:52 ERROR HTMLAnalyser:134 - Failed to canonicalise host: josep
hmoreau.schoolappointments.com: org.apache.commons.httpclient.URIException: gnu.inet.encoding.IDNAException: Contains non-LDH characters. josep%20hmoreau.schoolappointments.com
2017-02-21 22:21:28 INFO Instrument:155 - Performance statistics
WARCIndexerCommand.main#total(#=0, time=0.00ms, avg=0.00#/ms 0.00ms/#, 0.00%)
WARCIndexerCommand.parseWarcFiles#docdelivery(#=2089123, time=1706007.36ms, avg=1.22#/ms 0.82ms/#, 2.07%)
WARCIndexerCommand.checkSubmission#solradd(#=41782, time=1701567.49ms, avg=0.02#/ms 40.72ms/#, 2.06%)
WARCIndexerCommand.parseWarcFiles#startup(#=1, time=3372.23ms, avg=0.00#/ms 3372.23ms/#, 0.00%)
WARCIndexerCommand.commit#success(#=30, time=2663523.35ms, avg=0.00#/ms 88784.11ms/#, 3.23%)
WARCIndexerCommand.parseWarcFiles#fullarcprocess(#=30, time=1445718780.22ms, avg=0.00#/ms 48190626.01ms/#, 1754.28%)
WARCIndexerCommand.parseWarcFiles#solrdocCreation(#=8651855, time=77583908.73ms, avg=0.11#/ms 8.97ms/#, 94.14%)
WARCIndexer.extract#total(#=1740156, time=74867498.43ms, avg=0.02#/ms 43.02ms/#, 90.85%)
TextAnalyzers#total(#=1740156, time=5724356.71ms, avg=0.30#/ms 3.29ms/#, 6.95%)
PostcodeAnalyzer(#=1701619, time=46774.75ms, avg=36.38#/ms 0.03ms/#, 0.06%)
LanguageAnalyzer#total(#=1701619, time=5334955.20ms, avg=0.32#/ms 3.14ms/#, 6.47%)
LanguageDetector#startup(#=1, time=385.84ms, avg=0.00#/ms 385.84ms/#, 0.00%)
LanguageDetector.detectLanguage#li(#=1701619, time=2859452.22ms, avg=0.60#/ms 1.68ms/#, 3.47%)
LanguageIdentifier.addProfile(#=28, time=29.13ms, avg=0.96#/ms 1.04ms/#, 0.00%)
LanguageIdentifier#matchlanguageprofile(#=1701619, time=2551253.69ms, avg=0.67#/ms 1.50ms/#, 3.10%)
LanguageProfile.distanceInterleaved#total(#=47645332, time=2539920.08ms, avg=18.76#/ms 0.05ms/#, 3.08%)
LanguageProfile.Interleaved.update(#=1701647, time=216333.19ms, avg=7.87#/ms 0.13ms/#, 0.26%)
LanguageProfile.distanceInterleaved#dist(#=47645332, time=2236326.36ms, avg=21.31#/ms 0.05ms/#, 2.71%)
LanguageProfile#profilewriter(#=1701619, time=306254.26ms, avg=5.56#/ms 0.18ms/#, 0.37%)
LanguageDetector.detectLanguage#ld(#=1701562, time=2466748.82ms, avg=0.69#/ms 1.45ms/#, 2.99%)
FuzzyHashAnalyzer(#=1701619, time=337414.28ms, avg=5.04#/ms 0.20ms/#, 0.41%)
WARCIndexer.extract#hashstreamwrap(#=2089123, time=3746222.06ms, avg=0.56#/ms 1.79ms/#, 4.55%)
WARCIndexer.extract#analyzetikainput(#=1740156, time=64287494.76ms, avg=0.03#/ms 36.94ms/#, 78.01%)
WARCPayloadAnalyzers.analyze#total(#=1740156, time=64284418.35ms, avg=0.03#/ms 36.94ms/#, 78.00%)
WARCPayloadAnalyzers.analyze#firstbytes(#=1740156, time=17903.05ms, avg=97.20#/ms 0.01ms/#, 0.02%)
ARCNameAnalyzer.analyze(#=1740156, time=13765.56ms, avg=126.41#/ms 0.01ms/#, 0.02%)
XMLAnalyzer.analyze(#=7975, time=7579.49ms, avg=1.05#/ms 0.95ms/#, 0.01%)
WARCPayloadAnalyzers.analyze#arcname(#=1740156, time=14751.68ms, avg=117.96#/ms 0.01ms/#, 0.02%)
WARCPayloadAnalyzers.analyze#tikasolrextract(#=1740156, time=33048823.86ms, avg=0.05#/ms 18.99ms/#, 40.10%)
TikaExtractor.extract#detect(#=1740156, time=8280514.40ms, avg=0.21#/ms 4.76ms/#, 10.05%)
TikaExtractor.extract#extract(#=1735438, time=616945.84ms, avg=2.81#/ms 0.36ms/#, 0.75%)
TikaExtractor.extract#parse(#=1735438, time=23454988.10ms, avg=0.07#/ms 13.52ms/#, 28.46%)
ImageAnalyzer.analyze#facesanddominant(#=2547, time=1933767.38ms, avg=0.00#/ms 759.23ms/#, 2.35%)
WARCPayloadAnalyzers.analyze#droid(#=1740156, time=10547860.60ms, avg=0.16#/ms 6.06ms/#, 12.80%)
HTMLAnalyzer.analyze#total(#=1683855, time=18072449.77ms, avg=0.09#/ms 10.73ms/#, 21.93%)
HTMLAnalyzer.analyze#parser(#=1683855, time=12592642.58ms, avg=0.13#/ms 7.48ms/#, 15.28%)
HtmlFeatureParser.parse#jsoup(#=1683855, time=9542197.47ms, avg=0.18#/ms 5.67ms/#, 11.58%)
HtmlFeatureParser.parse#featureextract(#=1683855, time=2142947.86ms, avg=0.79#/ms 1.27ms/#, 2.60%)
PDFAnalyzer.analyze(#=5926, time=636542.70ms, avg=0.01#/ms 107.42ms/#, 0.77%)
WARCIndexer.extract#arheaders(#=2882905, time=693994.36ms, avg=4.15#/ms 0.24ms/#, 0.84%)
SolrRecord.removeControlCharacters#total(#=235107394, time=1551471.12ms, avg=151.54#/ms 0.01ms/#, 1.88%)
SolrRecord.sanitiseUTF8(#=235107394, time=469633.81ms, avg=500.62#/ms 0.00ms/#, 0.57%)
Parsing Archive File [31/131]:/data/ALBERTA_lfrancophonie_de_louest_canadien/ARCHIVEIT-3835-SEMIANNUAL-22840-20141107010041205-00007-wbgrp-crawl104.us.archive.org-6446.warc.gz
```

100 Million Solr Docs

525GB Solr index

Welcome to the Web Archives for Longitudinal Knowledge (WALK) portal. Before diving in, we encourage you to visit our [about](#) page.

Search

Advanced Search

General Content Type 3

- + html 716
- + other 72
- + pdf 10

Crawl Years 10

- + 2014 585
- + 2013 66
- + 2015 37
- + 2007 31
- + 2008 28
- + 2009 25
- + 2012 22
- + 2011 19
- + 2010 14
- + 2006 3

Links to Public Suffixes 10

- + com 704
- + ca 689
- + org 570
- + tv 528
- + de 408
- + net 282
- + gov 106
- + it 104

Sample Mode rob ford + crack

Search

Reset

• Search Term(s): rob ford + crack

Results

Concordance

Results 1 to 10 of 798

CSV

Asc

« 1 2 3 4 5 6 »

Action

1 "Newspaper Articles Directory - 2006, news, child rights, child support, children, divorce, family law, adoption, Toronto Star, National Post, Globe and Mail, Sun"
Oct 20 2006 09:48:02 EDT
html
www.canadiancrc.com

Shine

WARClight prototype

The screenshot shows the WARClight web application interface. The browser's address bar displays the URL `localhost:3000/?utf8=✓&search_field=all_fields&q="Nick+Ruest"`. The application header features the "WALK" logo and navigation links for "Apps", "GMail", "Lib", "GitHub", "AWS", "HistD", "RSS", "Globe", "WALK", "LEARN", "FT", "Blacklight", "Blacklight repo", and "Concur". A search bar at the top contains the query "Nick Ruest" and a "Search" button. Below the search bar, a "Limit your search" section allows filtering by "General Content Type" (html: 47, other: 2, pdf: 1, text: 1, audio: 0, excel: 0, image: 0, powerpoint: 0, video: 0, word: 0) and other criteria like Domain, Links Domains, Institution, Collection Name, and Collection Number. The search results are displayed in a list, with the first two items visible:

- 1. 트위터의 nick ruest (ruebot)님**
Title: 트위터의 nick ruest (ruebot)님
Host: twitter.com
Crawl Date: 2013-10-11T00:58:16Z
Content Type: text/html
Domain: twitter.com
This page links to: twitter.com, t.co, and twimg.com
Institution: University of Alberta Libraries
Collection Name: Alberta Floods June 2013
Collection Number: 3747
- 2. AhemNason (Mike Nason) / followers · GitHub**
Title: AhemNason (Mike Nason) / followers · GitHub
Host: github.com
Crawl Date: 2016-07-26T08:06:00Z
Content Type: text/html
Domain: github.com
This page links to: github.com
Institution: Simon Fraser University Library
Collection Name: Public Knowledge Project
Collection Number: 7100

The interface also includes pagination controls ("« Previous | 1 - 10 of 51 | Next »") and a "10 per page" selector. A "Bookmark" checkbox is visible next to each result item.

Finding famous people in WARClight, by Ian Millign



Social Sciences and Humanities
Research Council of Canada

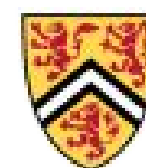
Conseil de recherches en
sciences humaines du Canada

Canada



compute
canada

calcul
canada



UNIVERSITY OF
WATERLOO

Thanks very much!

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Nick Ruest
Digital Assets Librarian
@ruebot

